Ch/S of big data?

① Data Volume :-

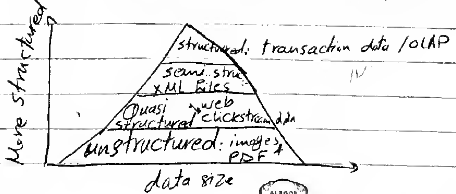
→ 44x increase from 2009 to 2020
(0.8 Zettabytes to 35.2 Zb)
So, highly rate of growth

② processing Complexity :-

- → Changing data structures
- → use cases that require additional transformations & different analytical techniques.
- → processing big data is in parallel Computing enviroments and parallel processing, wich enable parallel loading, and analysis of data

③ Data structure :-

- → Greater Variety of data structures to mine and analyze
- → Most of the big data is unstructured or Semi-structured, Wich requires different techniques and tools to process & analyze.


More Structured
structured: transaction data / OLAP
semi-struc
XML files
Quasi structured
web clickstream data
unstructured: images + PDF
data size

Business intelligence Vs. Data sceince

1 Data sceince (predictive Analytics & Data Mining)

A. Typical techniques & Data types:-

- [optimization, predictive modeling] [statistical analysis], [machine learning techniques as Naive Bayes or regression]

- Structured / unstructured data. ~~many~~ many types of sources.
  Very large data sets.

B. Common Questions :-

1) what if 2) what's the optimal scenario for or business?

3) What'll happen next? 4) Why is this happening?

"open ended questions"

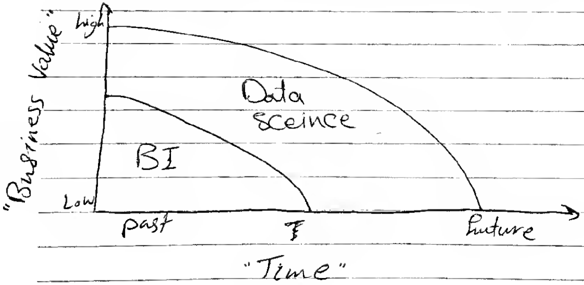
2) Business intelligenee: (BI)

A_ techniques & Data types :

- Standard and ad hoc reporting, alerts, queries, and details on demand
- Structured data, traditional sources, and mageable data sets.

B. Common Questions :-

- what happend last quarter?
- Holi many did we sell
- where is the problem? in wich situations?



Big data :-

data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new source of business value.

So it requires

1) new data architectures, analytic sandbox
2) new tools to store, mange and realize the business benefit of these large data
3) new analytical methods

another definition of big data

1) Volume 2) velocity 3) Varaity
4) Value 5) Veracity

-> Volume : Lagre increase of data volume [illegible]

1) The transactional data that has added up over years
2) Straming data from social media
3) Machine to machine data [illegible]

-> velocity : Data is being streamed at huge speed and needs to be dealt with a timely manner
- -> Social Media
- -> Mobile Devices

-> Varaity : because nowadays we deal with different kinds of data structures

1) Structured 2) [illegible]meric 3) unstructured

4) Email 4) Audi[illegible]

Data repositories

1) Data islands → Spreadsheets and low volume DB just for record
→ Analyst dependent on data extraction

2) Data warehouses → Support BI reporting, but restricts robust analyses or data exploration
→ Analyst dependent on IT & DBAs for data access and schema changes
→ Data takes big time to get extract from multiple sources.

3) Analytic Sandbox :

→ Enables high performance analytics using in-db processing

→ reduces costs associated with data replication into "shadow" file systems

→ analyst owned not DBA owned

→ More robust analyst.

Subject: lec 2 **Date:**

profile of data scientist

1) Quantitative [illegible]
2) Technical
3) Curious & Creative
④ Skeptical
5) Communicative & Collabrative. [illegible]

lecture 3

Data Analytics Lifecycle

① value of using it ?
→ focus your time
→ Ensure rigidity & completeness
→ Enable better to members of the cross-functional analytic teams

② Key Roles for a successful Analytic project..

→ Business user → someone who benefits from the end result

→ project sponsor:. provides the funding & will measure the degree of value from the final outputs

→ project Manager:. Ensure key objectives are met on time with good quality

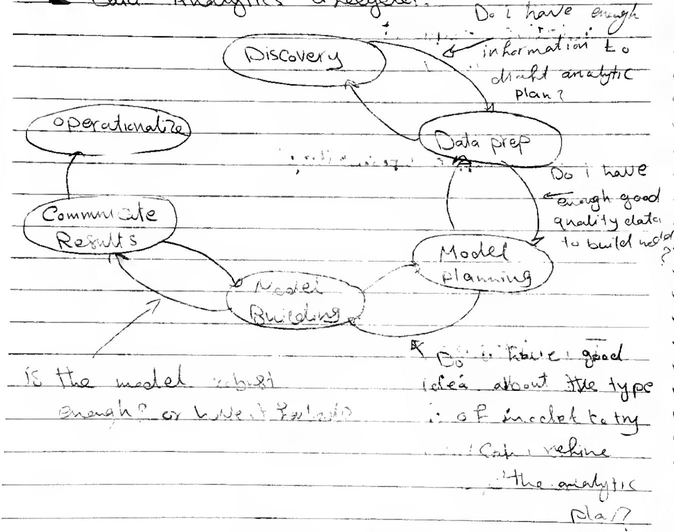
→ Business intelligence Analyst:. expertise with deep understanding of the data, and business intelligence from reporting perspective

→ Data Engineer:. assist with tuning sql queries for data management

→ Database Administrator:. configures database environment to support the analytical needs

Data scientist :- provides subject matter expertise for analytical techiques, data modeling, applying valid analytical techuques to given business problem

- Data Analytics Lifecycle.


Discovery
Do I have enough information to draft analytic plan?
Data prep
Do I have enough good quality data to build model?
Model planning
Do I have good idea about the type of model to try
Can I refine the analytic plan?
Model Building
Communicate Results
Is the model robust enough? or have I failed?
Operationalize

المحاضرة الرابعة [illegible]

association rules :- another supervised learning method it is used to discover relations within the data

→ Not predictable method used to find similarity

Subject: ______________ Date: ______________

① association Rules: step ① "A priori Algorithm"

absolute Support Count: عدد تكرار ال X داخل ال data ;X وتعريفه

Frequency or occurrence of an itmeset X

relative Support the fraction of transaction that contains X (probability that a transaction contains)

$\frac{X_i}{\sum X}$

Step ②

S → Support (probability) طالعين مع X, Y كام مرة

Confidence → Conditional probability that transaction having X also contains Y

Example

Tid	itemes
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Sup = 2

itemset	Sup
A	2
B	3
C	3
D	1
{E}	3

itemset	Sup
{A}	2
{B}	3
{C}	3
{E}	3

itemset	Sup
~~{A,B}~~	1
{A,C}	2
~~{A,E}~~	1
{B,C}	2
{B,E}	3
{C,E}	2

itemset	Sup
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

itemset	Sup
{B,C,E}	2

(2) ANOVA F test

$K = 3$ = No. populations

D_1	D_2	D_3
1	2	2
2	4	3
5	2	4

N = No. of all numbers (samples)

(1) mean of each population

$m_1 = \frac{1}{3}(1+2+5) = 2.67$

$m_2 = \frac{1}{3}(2+4+2) = 2.67$

$m_3 = \frac{1}{3}(2+3+4) = 3$

(2) grand mean (m_0)

$= \frac{1}{3}(2.67+2.67+3) = 2.78$

(3) $SS_{within} = \sum_i (x_{1i}-m_1)^2 + \sum_i (x_{2i}-m_2)^2 + \sum_i (x_{3i}+m_3)^2$

(4) $SS_{total} = \sum_i (x_i - m_0)^2$

(5) $SS_{between} = SS_{total} - SS_{within}$

(6) $S_W^2 = \frac{SS_{within}}{N-K}$

(7) $S_B^2 = \frac{SS_{between}}{K-1}$ $F = \frac{S_B^2}{S_W^2}$

$F > 1$ difference between hypothesis is big reject Null.

$F < 1$ don't reject the Null.

ALAQSA

Subject: Lec 11 **Date:**

Naive Bayesian Classifiers

determine the most probable class label for each object.

بس بيتعامل على أساس ان كل عنصر مش معتمد على التاني مثلاً

given object {Shape, Color, weight}

وحتى لو كان في أي عنصر معتمد ع التاني هيعمل عيم وهيفرض ان ملهمش علاقة ببعض.

input → discrete

output ① probability score
② class label

example for application Spam filtering

practical example..

A(male single, skilled, own, <100)

if P(good | A) > P(bad | A) A is good

والعكس

P(A|good) = P(male single | good) * P(skilled | good)
* (own | good) * P(<100 | good)

$\frac{14}{19} * \frac{12}{19} * \frac{13}{19} * \frac{10}{19} =$

$P(good) = \frac{19}{27}$ = ...

$P(good|A) = \frac{P(A|good) * P(good)}{P(A)}$

والعكس لا good

Subject: Date:

ازاي أعرف الـ model اللي اخترته ده بيطلع نتائج بدقة كويسة
و الا لا : هنعمل خليه Diagnostics وفيه
1) Confusion Matrix
2) RoC Curve & AUC ، ص 16

Confusion Matrix

	good	bad
	TP	FN
	FP	TN

هنقول انه good ← P
N ← bad

TP : يعني ادى انه good وهو فعلا good

FN يعني قال انه bad وهو فعلا good

FP قال انه good وهو bad

TN قال انه bad وهو bad

① Success → accuracy $\frac{\Sigma T}{\Sigma \text{كله}} = \frac{TP+TN}{TP+TN+FP+FN}$

② "Recall" or TPR (True positive Rate) $\frac{TP}{TP+FN}$

③ FPR (False Positive Rate) $\frac{FP}{FP+TN}$

④ FNR (False -ve Rate) $= \frac{FN}{TP+FN}$

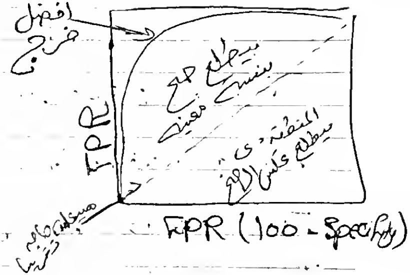
5) precision :- نسبة الحاجة الي قالها +ve و P فعلا +ve
$TP/(TP+FP)$

Subject: محاضرة م١٢ Date:

of Naive Bayesian

(+)	(-ve)
handles missing Values quiet Well	

2 Roc Curve



Subject: Date:

lecture 10 | logistic regression

[1] بيطلع probability / وممكن يشتغل كـ classifier

انت هتاخد ال probability بنسبة كذا تحطها في classifier معين

(2) input → Continous "float" or → discrete "1,2,3.."

(3) output → prediction
could be (Binary classifier case).

odds ratio = $\frac{P(x=true)}{P(x=false)}$

[4] When to use logistic regression

[1] preferred method for binary classification problems:
وخاصة لو انت محتاج تطلع probability من الموديل
بجانب تصنيف ال data متاكد بتبانيه لومنفعش شوق غير

(2) examples :- probability that a customer will churn or not
this is binary classifier 0, 1

(3) multi class example:-
0, 1, 2

(15/15) 3c

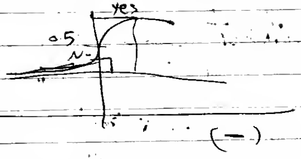
خلاصة المثال هو ✓

ده رسم الداتا أو الرسم اللي هو عبارة عن تصنيف

ال borrowers يا إما يا آه أو لا وهيتم بين الناس

اللي الاحتمالية بتاعتهم أكتر من اه

slide 7 لو طلعنا probability خلاص وعايز تصنيف نعمل



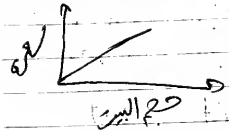
= threshold

هنذاكره

---

lec : 9

Linear regression = relationship between an outcome and its inputs variables



مثلاً

use cases :

① estimate lifetime value of customers

② estimate the probability that loan will default or not